PATENT
ATTY. DOCKET NO. IBM/156

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

Ex parte Abdo et al.

Appeal No. ________

Serial No.:	**10/ 017,783**
Filed:	**December 13, 2001**
Group Art Unit:	**2177**
Examiner:	**Mohammad Ali**
Applicant:	**Abdo et al.**
Title:	**ESTIMATION AND USE OF ACCESS PLAN STATISTICS**

Cincinnati, Ohio 45202

April 3, 2006
Via EFS-WEB

APPEAL BRIEF

This brief is in furtherance of Applicant's Notice of Appeal filed February 1, 2006, appealing the decision of the Examiner dated November 1, 2005 finally rejecting claims 1-10. A copy of the claims appears in the Appendix to this brief.

Certificate of Electronic Transmission

I hereby certify that this correspondence and any enclosures are being electronically transmitted via EFS-WEB on the date indicated below.

/Thomas W. Humphrey/ April 3, 2006
Thomas W. Humphrey Date
Reg. No. 34,353

Real Party In Interest

The real party in interest in this appeal is INTERNATIONAL BUSINESS MACHINES CORPORATION, a corporation of New York having a place of business at New Orchard Road, Armonk, New York 10504.

Related Appeals and Interferences

There are no such appeals or interferences.

Status of Claims

Claims 1-10, stand rejected under 35 U.S.C. §103, asserted to be obvious over Chada, U.S. Patent 5,706,495 ("Chadha '495") in view of Jones et al., U.S. Patent 5,689,698 ("Jones") and Chadha et al., U.S. Patent 6,032,146 ("Chadha '146").

Claims 1-24 were originally filed with the application. Applicant's Response to Office Action dated July 9, 2004 did not change claims 1-24. Applicant's Response to Restriction Requirement dated February 25, 2005 elected claims 1-10 and withdrew claims 11-24. Applicant's Response to Office Action dated August 11, 2005, amended Claim 1 only.

Status of Amendments

There are no amendments pending.

Summary of Claimed Subject Matter

The claims of this application relate to the use of indexes and statistics in forming access plans for a database. An index is frequently used to access desired data in a database, and often statistics are used prior to executing a query, to estimate the likely size of the solution sets that will be generated from each selection criterion in a query. Typically, statistics are generated using an index in combination with the specific selection criterion of the query being processed.

For example, as discussed in the specification at pages 1-2, consider a car owners database, and a query seeking persons named "Smith" that live in the city of "New York" and own a "Packard" car. For this query, statistics for each of these criteria would be generated from an index, i.e., how many persons in the database are named "Smith", how many live in "New York", and how many own a "Packard" car. These statistics would likely show that the most efficient path to the desired result would be first seeking "Packard" owners, then selecting those in "New York" named "Smith". This would be far more efficient than first identifying "New York" city residents or persons named "Smith". The statistics, generated from the index, would thus identify the most efficient path to the answer.

As noted in the specification at page 3, after collecting statistics for a given query, those statistics may be cached for later re-use. For example, subsequent queries that seek persons in the city of "New York", could re-use the statistic previously generated for that same criterion. However, in order to re-use statistics, those statistics must be valid for the current criterion. The number of rows having a city name of "Brainerd" might be substantially less than the number having a city name of "New York". Thus, statistics generated for a first selection criterion like "New York" on a given attribute, typically cannot be reused for a second selection criterion such as "Brainerd", but rather must be re-validated by re-accessing the associated index.

Unfortunately, the time required to access indexes to generate statistics can be a substantial fraction of total query optimization time; thus, re-validation of statistics represents a substantial loss of efficiency in database processing.

As summarized at page 6 of the specification, claims 1-10 relate to the use of previously generated statistics generated for an attribute, on different queries. In processing a query including a selection criterion on one or more attributes of a relation, a prior statistic generated for a prior different selection criterion on the same one or more attributes of the relation, may be used in processing the query, even though the selection criterion differs. According to the claimed invention, the decision to re-use a criterion, is based upon a measure of the entropy of the one or more attributes of the relation. Specifically, if the entropy measure suggests that different attribute values generate similar numbers of hits, then the statistics need not be re-computed. In this way, the re-validation of statistics may be performed more efficiently.

Page 29 of the specification and Fig. 5 illustrate the process for validating statistics based upon entropy data. Specifically, a threshold value is selected and it is determined whether entropy values are greater than the threshold.

An example of an entropy measure is the frequency with which values recur in a particular attribute, which may be expressed relative to the total number of distinct values as a probability.

Claims 1, 7 and 8 of the present application references "revalidating" prior or previously generated statistics for a query by "identifying in said query a selection criterion" and "revalidating a prior statistic generated for a prior different selection criterion on the same one or more attributes of [the] relation, based upon a measure of entropy of said one or more attributes of said relation". Dependent claims provide specifics for this concept, such as claim 2 which states the use of a "threshold value" for revalidating a "measure of entropy", and claims 3-6 which states a measure of entry is generated by "computing frequencies of different values for the one or more attributes" and "combining the measured frequencies".

Grounds of Rejection

In rejecting claims 1-10, the Examiner has relied upon the Chadha '495, Chadha ‘146 and Jones patents. The Examiner appears to be asserting that Chadha ‘495 teaches holding of statistics on data to be accessed “(the size of the table, the number of distinct values in a particular column, etc.)” and using statistics to choose an efficient access path. The Examiner concedes that Chadha ‘495 does not teach revalidating statistics, but asserts that Jones teaches “revalidation” and references col. 15, lines 10-15 and Fig. 40 of Jones. The Examiner concedes that neither Chadha ‘495 nor Jones disclose “entropy”, but asserts that Chadha ‘146 teaches entropy “(rules used for data mining will appreciate that the association measures can be Chi-square, entropy, see col. 5, lines 14-15.)”. The Examiner asserts that an obvious combination of Chadha ‘495, Jones and Chadha ‘146 would lead to the claimed invention.

Argument

Chadha '495, relied upon by the Examiner for the use of statistics, generally discloses the use of a particular index type, known as the encoded vector index, to index the content of a relation. The Examiner cites to text in column 4 of Chadha '495 describing an attribute as a field or column of a relational database, and to text in column 7 of Chadha '495 stating that statistics are generated from an index to optimize an access plan, and to text in column 9 stating how updating of an encoded vector index is handled when new tuples are inserted into a relation.

The disclosure of Chadha '495 thus generally includes the notion of forming "statistics" on a relation, and explains its relevance to an encoded vector index. However, Chadha '495 does not deal with the issues that are at the heart of the claimed invention. Specifically, Applicant has found nothing in Chadha '495 that in any way relates to the claimed steps of computing "a measure of entropy of attribute values" and/or "revalidating a prior statistic." There is simply nothing in Chadha '495 that describes the revalidation of statistics, i.e., determining whether those statistics are applicable to a new and different query than the query for which they were formed. Furthermore, nothing in Chadha '495 discusses the use of correlation values between attributes or other "entropy" figures to validate statistics. Rather, Chada '495 simply generates statistics (which may be the encoded vector index or something derived from it) and uses the statistics to form an access plan, as is known from the prior art acknowledged in the present application.

The Examiner's citation of the Jones patent is also inappropriate to the claimed invention. Jones relates to accessing a database in response to a query. The text noted by the

Examiner – column 7 of Jones – states that statistical information is used to optimize a query plan. As such, Jones is no different than the background art noted above.

The Examiner has noted text in column 15 of Jones, which states that when data is to be sent to a receiver from the object server, the first step of performing this transmission is the "revalidation" of the request, followed by the formulation of a query plan. The Examiner appears to believe that the "revalidation" identified in column 15 is somehow descriptive of associating statistics for one selection criterion, to a second different criterion, as claimed. Applicant has been unable to find any support for such an interpretation of Jones. Rather, the "revalidation" described in column 15 appears to relate to the request for or transport of data that is to be sent to the receiver client, not to any statistics. Applicant notes that the term "validate" is used in Jones to refer to transport sessions, not to statistics, as seen for example at column 14, line 8. Furthermore, there is no apparent connection between the "revalidate" step 606 described in column 15, and the subsequent steps in which a "query plan is formulated. This query plan formulation process is analogous to that previously described." Rather, this text appears to indicate that query plan formulation is separate from the "revalidation" of the request, not part of it as the Examiner posits.

In short, Applicant finds nothing in Jones to suggest anything more than the use of statistics in the normal fashion, to create access plans. More particularly, there is nothing in Jones to suggest the claimed invention of "revalidating a prior statistic generated for a prior different selection criterion", by the use of a "measure of entropy of [the] one or more attributes" of a relation.

The Examiner’s reliance upon Chadha ‘146 is also misplaced. Chadha ‘146 discloses a data mining method in which attributes of a relation are combined, based upon various criteria (dimension reduction), so that data mining can be more efficiently performed. This method and analysis does relate – tangentially – to entropy and the correlation of attributes, as "data mining" may involve entropy computations, as the Examiner has noted in col. 5 of Chadha ‘146. However, in no way does the method or topic of Chadha ‘146 relate to the point of this application and its claims, which is the re-use of statistics created for one database access plan, in another database access plan. More specifically, Chadha ‘146 in no way relates to "revalidating a prior statistic generated for a prior different selection criterion", by the use of a "measure of entropy of [the] one or more attributes" of a relation.

The Examiner's citation to text in Chadha ‘146 relating to dimension reduction, such as in column 7, in no way relates to generation of statistics in accesses to a database or the re-use of such statistics in subsequent accesses of a database.

In summary, then, the Examiner’s rejection is an improper combination of references having nothing to do with the problem addressed by the present invention nor the solution presented by the claims. None of the references relates to “revalidating” statistics generated for one query, for use with another query, nor do they relate to the use of “entropy” values in making a “revalidation”. The Examiner appears to have selected the Jones and Chadha ‘146 references merely because they use the words “entropy” and “revalidate”, not because they have any contextual relationship to the claimed invention.

The Examiner's rejections of various dependent claims rely upon one or more of Chadha ‘495, Chadha ‘146 and Jones. For the reasons noted above, which need not be repeated, Applicant respectfully submits that none of these references are relevant to the claimed invention. Furthermore, the Examiner's remarks fail to identify any citation in any of the three references that shows a "threshold" being applied to an "entropy" for the purpose of revalidation, or show an "entropy" being computed for this use by determination of the number of times a value appears.

Applicant therefore respectfully submits that the Examiner's rejection is in error and a reversal of the rejection and allowance of the claims is therefore requested.

Respectfully submitted,
Wood, Herron & Evans, L.L.P.

/Thomas W. Humphrey/

Thomas W. Humphrey
Reg. No. 34,353

2700 Carew Tower
441 Vine Street
Cincinnati, OH 45202-2917

Voice: (513) 241-2324
Facsimile: (513) 241-6234

Claim Appendix

1. (Previously presented) A computer-implemented method for revalidating previously generated statistics for a query directed to one or more attributes of a relation, comprising

identifying in said query a selection criterion on said one or more attributes of said relation, and

revalidating a prior statistic generated for a prior different selection criterion on the same one or more attributes of said relation, based upon a measure of entropy of said one or more attributes of said relation.

2. (original) The method of claim 1 wherein said prior statistic is revalidated if a measure of entropy of said one or more attributes of said relation is less than a predetermined threshold value.

3. (original) The method of claim 1, further comprising generating a measure for the entropy of said one or more attributes of said relation, by the steps of

computing frequencies of different values for the one or more attributes in tuples of the relation, and

combining the measured frequencies into a measure of the entropy of the attributes.

4. (original) The method of claim 3, wherein generating a measure for the entropy of said one or more attributes of said relation further comprises collecting a sample of tuples of the relation, wherein frequencies of different values are computed for tuples in the sample.

5. (original) The method of claim 3 wherein combining the measured frequencies comprises determining a number of distinct values for the one or more attributes, and converting the computed frequencies to probabilities by dividing the frequencies by number of distinct values.

6. (original) The method of claim 5 wherein combining the measured frequencies further comprises forming a weighted sum of the computed probabilities.

7. (original) A computer system implementing a relational database system and evaluating queries directed to said relational database, comprising

storage for said relational database, including a relation having a plurality of tuples including values for a plurality of attributes, and

computing circuitry performing query optimization and query execution upon said relational database, said query optimization including generating statistics for a query directed to one or more attributes of said relation, by identifying in said query a selection criterion on said one or more attributes of said relation, by revalidating a prior statistic generated for a prior different selection criterion on the same one or more attributes of said relation, based upon a measure of entropy of said one or more attributes of said relation.

8. (original) A program product for implementing a relational database system and evaluating queries directed to said relational database, comprising

a relational database, including a relation having a plurality of tuples including values for a plurality of attributes, and

relational database software performing query optimization and query execution upon said relational database, said query optimization including generating statistics for a query directed to one or more attributes of said relation, by identifying in said query a selection criterion on said one or more attributes of said relation, by revalidating a prior statistic generated for a prior different selection criterion on the same one or more attributes of said relation, based upon a measure of entropy of said one or more attributes of said relation, and

a signal bearing media holding said relational database and relational database software.

9. (original) The program product of claim 8 wherein the signal bearing media comprises transmission media.

10. (original) The program product of claim 8 wherein the signal bearing media comprises recordable media.

11. (withdrawn) A method for identifying a group of attributes of a relation for which a multi-dimensional index is to be formed, comprising

computing a correlation of attribute values within tuples of the relation, and

forming a multi-dimensional index for a group of attributes within tuples of the relation having a correlation of attribute values in excess of a threshold.

12. (withdrawn) The method of claim 11, wherein computing a correlation of attribute values within tuples of the relation comprises collecting a sample of tuples of the relation, and computing correlation of attribute values within the sampled tuples.

13. (withdrawn) The method of claim 11 wherein a correlation of attribute values is computed as an information gain for those attributes by comparing, for a common set of tuples, a sum of individual entropies of values of each attribute, to a joint entropy of the values of all attributes.

14. (withdrawn) The method of claim 13, wherein a measure for the entropy of one or more attributes is generated by computing frequencies of different values for the one or more attributes in tuples of the relation, and combining the measured frequencies into a measure of the entropy of the one or more attributes.

15. (withdrawn) The method of claim 14, wherein generating a measure for the entropy of said one or more attributes of said relation further comprises collecting a sample of tuples of the relation, wherein frequencies of different values are computed for tuples in the sample.

16. (withdrawn) The method of claim 14 wherein combining the measured frequencies comprises determining a number of distinct values for the one or more attributes, and converting the computed frequencies to probabilities by dividing the frequencies by number of distinct values.

17. (withdrawn) The method of claim 16 wherein combining the measured frequencies further comprises forming a weighted sum of the computed probabilities.

18. (withdrawn) The method of claim 11 further comprising evaluating attribute groups found to have correlation to identify primary sources of correlation, by determining a mutual information gain by comparing information gain for a group of attributes, to the largest information gain of any sub-group of fewer of the same attributes.

19. (withdrawn) The method of claim 18 wherein a multi-dimensional index is formed for an attribute group having information gain greater than a

threshold, if there is no larger attribute group including the same attributes having a mutual information gain greater than a threshold.

20. (withdrawn) The method of claim 11 wherein correlation of attribute values is computed for all combinations of attributes of a relation, or alternatively by sampling a set of attribute groups and then evaluating other related groups of those found to have substantial correlation.

21. (withdrawn) A computer system implementing a relational database system including indexes for said relational database, comprising

storage for said relational database, including a relation having a plurality of tuples including values for a plurality of attributes, and

computing circuitry performing query execution upon said relational database, and identifying a group of attributes of a relation for which a multi-dimensional index is to be formed, by computing a correlation of attribute values within tuples of the relation, and forming a multi-dimensional index for a group of attributes within tuples of the relation having a correlation of attribute values in excess of a threshold.

22. (withdrawn) A program product for implementing a relational database system, comprising

a relational database, including a relation having a plurality of tuples including values for a plurality of attributes, relational database software performing query execution upon said relational database, and identifying a group of attributes of a relation for which a multi-dimensional index is to be formed, by computing a correlation of attribute values within tuples of the relation, and forming a multi-dimensional index for a group of attributes within tuples of the relation having a correlation of attribute values in excess of a threshold, and

a signal bearing media holding said relational database and relational database software.

23. (withdrawn) The program product of claim 22 wherein the signal bearing media comprises transmission media.

24. (withdrawn) The program product of claim 22 wherein the signal bearing media comprises recordable media.

Evidence Appendix

None.

Related Proceedings Appendix

None.

TABLE OF CONTENTS